

Contents lists available at [ScienceDirect](http://www.sciencedirect.com)

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

Mining association language patterns using a distributional semantic model for negative life event classification

Liang-Chih Yu ^{a,*}, Chien-Lung Chan ^a, Chao-Cheng Lin ^b, I-Chun Lin ^{c,d}^a Department of Information Management, Yuan Ze University, Chung-Li, Taiwan, ROC^b Department of Psychiatry, National Taiwan University Hospital and National Taiwan University College of Medicine, Taipei, Taiwan, ROC^c Department of Computer Science and Information Management, Hungkuang University, Taichung, Taiwan, ROC^d Department of Industrial Management, National Yunlin University of Science and Technology, Yunlin, Taiwan, Republic of China

ARTICLE INFO

Article history:

Received 17 November 2009

Available online 1 February 2011

Keywords:

Text mining

Natural language processing

Association language pattern

Distributional semantic model

Mental health

Negative life event

ABSTRACT

Purpose: Negative life events, such as the death of a family member, an argument with a spouse or the loss of a job, play an important role in triggering depressive episodes. Therefore, it is worthwhile to develop psychiatric services that can automatically identify such events. This study describes the use of association language patterns, i.e., meaningful combinations of words (e.g., <loss, job>), as features to classify sentences with negative life events into predefined categories (e.g., Family, Love, Work).

Methods: This study proposes a framework that combines a supervised data mining algorithm and an unsupervised distributional semantic model to discover association language patterns. The data mining algorithm, called association rule mining, was used to generate a set of seed patterns by incrementally associating frequently co-occurring words from a small corpus of sentences labeled with negative life events. The distributional semantic model was then used to discover more patterns similar to the seed patterns from a large, unlabeled web corpus.

Results: The experimental results showed that association language patterns were significant features for negative life event classification. Additionally, the unsupervised distributional semantic model was not only able to improve the level of performance but also to reduce the reliance of the classification process on the availability of a large, labeled corpus.

© 2011 Elsevier Inc. All rights reserved.

1. Introduction

In their daily lives, people may suffer from stressful or negative life events such as the death of a family member, an argument with a spouse or the loss of a job. Such negative life events have been recognized as being associated with the onset of depressive episodes [7,41]. Therefore, many psychiatric websites have been developed for mental health care and prevention. The representative websites include Depression Forum,¹ WebMD,² SA-UK,³ Yahoo! Answers,⁴ John Tung Foundation⁵ and PsychPark⁶ [3,28]. These websites provide community-based services for Internet users to share their life stresses and depressive problems with other users and health professionals. That is, users can describe their stressful or

negative life events along with any depressive symptoms via web forums, message boards, and blogs. Other users or health professionals can then make recommendations in response to these problems. Fig. 1 shows an example of a forum post and responses to the post.

The major characteristic of forum posts is that they contain large amounts of natural language expressions related to negative life events, which are useful for developing language systems that can automatically identify negative life events. Additionally, knowing about negative life events can make online psychiatric services more effective. For instance, an ideal psychiatric retrieval system should be able to retrieve relevant forum posts according to the negative life events experienced by users so that the users can see that they are not alone—many people have suffered from the same or similar problems. They can then learn self-help knowledge from the responses. Consider the example post in Fig. 1 as an input query. This example indicates that a user experienced a socially related negative life event. Therefore, the retrieval system could rank the posts containing social events higher than those containing other events (e.g., family events) if it could correctly identify the negative life event described in the example post. Additionally, a text summarization system capable of analyzing negative life events from forum posts could assist health professionals capture

* Corresponding author. Address: 135 Yuan-Tung Road, Chung-Li 32003, Taiwan, ROC. Fax: +886 3 435 2077.

E-mail address: lcyu@saturn.yzu.edu.tw (L.-C. Yu).

¹ <http://www.depressionforums.org/forums/forums.html>.

² <http://www.webmd.com>.

³ <http://www.social-anxiety-community.org/db>.

⁴ <http://answers.yahoo.com>.

⁵ <http://www.jtf.org.tw>.

⁶ <http://www.psychpark.org>.

Post:

I don't have a lot of friends. Last week, I argued with my best friend and was upset. So I have no friend to hang out with. My life now involves working all day, then coming home to watch TV and play Internet games. Everyday repeats the same routine. I feel so alone and hate life like this.

Response 1:

I'm sorry that you are so alone. Maybe you can join online communities like Meetup.com. They have meetups for people who have same hobbies.

Response 2:

I know some ways you can make new friends – do some volunteer work, join church groups if you are religious, look for a team to play on if you like sports, take classes you like in the weekend. I hope you feel better. Good luck.

Fig. 1. Example of forum posts and responses.

users' background information more quickly and thereby speed up the response time. A dialog system could generate supportive responses if it could understand the negative life events embedded in users' input. Therefore, this study proposes a framework for negative life event identification. We formulate this problem as a sentence classification task; that is, classify sentences according to the type of negative life events within them. The class labels used herein are presented in Table 1 and were derived from recent studies on negative life events [7,39,55].

Traditional approaches to sentence classification [21,36] or text categorization [25,27,45] usually adopt the bag-of-words model as the baseline feature to train classifiers. For example, the bag-of-words can be used to train a naïve Bayes classifier by assuming that each word in the word bag is independent [21,27]. However, the independence assumption ignores the association between words in sentences, which may impose a limitation on classification performance. Therefore, extended Bayes classifiers such as associative naïve Bayes classifier [22], tree augmented naïve Bayes classifier [15], and semi-naïve Bayesian classifier [23] have been developed to improve the naïve Bayes classifier by relaxing the restrictive independence assumption. Another method to consider is the use of n -grams to capture sequential relations between words to boost classification performance [10,40,48,54]. The use of n -grams is effective in capturing the local dependencies of words, but tends to suffer from data sparseness problems in capturing long-distance dependencies because higher-order n -grams require large training datasets to obtain reliable levels of estimation.

For our task, there exist several one-to-many relationships in different layers in texts. For example, a corpus can have multiple sentences, and a sentence can have multiple words. The unit of analysis herein is a sentence. That is, this study performs a sentence-level analysis to extract the *association language patterns* [9] from the sentences in a corpus. An association language patterns represents a meaningful combination of words, such as <worried, children's, health>, <broke up, boyfriend>, <school, teacher, blames>, <lost, job>, and <argued, friend> in the example sentences in Table 1. These patterns can capture the dependencies of multiple words in the sentences, which help to understand the negative life event embedded in the sentences. Additionally, such patterns are not necessarily composed of continuous words. Instead, they are

usually composed of words with long-distance dependencies, which cannot be easily captured by n -grams. Therefore, this study proposes the use of association language patterns as features to build classifiers for negative life event classification.

In the acquisition of association language patterns, there are two main research approaches: knowledge-based [26,35] and corpus-based approaches [2,9,19,34,52]. Knowledge-based approaches rely on exploiting expert knowledge to design handcrafted patterns. A major limitation of such approaches is the requirement of significant time and effort to design the handcrafted patterns. Additionally, these patterns have to be redesigned when they are applied to a new domain. Such limitations form a knowledge acquisition bottleneck. Corpus-based approaches can discover language patterns on domain corpora using supervised learning techniques. The corpora must be labeled with domain-specific knowledge (e.g., events). Various statistical methods can then be adopted to discover language patterns from possible combinations of words in the corpora. For instance, *association rule mining* [1,11,13,46], which has been extensively studied in the data mining community, can be transformed to generate association language patterns from a corpus of sentences labeled with negative life events. However, supervised learning approaches require large amounts of labeled corpora, which are not easily obtained for some application domains. Accordingly, there is an emerging demand for a framework capable of learning from unlabeled corpora.

Therefore, semi-supervised learning techniques that can use both labeled and unlabeled data have been widely investigated in many fields [8,57]. These techniques include self-training [31,44], co-training [5,56], expectation maximization (EM) based methods [6,37,38], transductive support vector machine (TSVM) [20,33,49], and graph-based methods [47,58]. In self-training, a base classifier is first trained on the full feature set of the labeled data. The base classifier is then used to classify a portion of the unlabeled data, and the most confidently classified data examples are added to the labeled data. The base classifier is re-trained on the augmented labeled data, and the process is iterated. Contrary to the single view (feature set) used for building self-training classifiers, co-training assumes that the features of data examples can be partitioned into two different views (feature sets). Two distinct classifiers can then be trained on the two different views of the input labeled data, respectively. Each classifier then iteratively classifies the unlabeled data to augment the labeled data. In addition to using self-training and co-training alone, these two methods can be combined to develop a self-combined algorithm [16]. Additionally, co-training can also be combined with the EM algorithm to develop a Co-EM algorithm, which can be applied to both generative models such as naïve Bayes classifiers [37] and discriminative models such as SVMs [6]. Transductive VSM aims to find a better separating hyperplane using both labeled and unlabeled data. Graph-based methods construct a graph whose nodes denote labeled and unlabeled data examples and edges denote similarities between the data examples. Various methods can then be used to propagate the label information from the labeled examples to the unlabeled examples. Recently, a unified theoretical framework for semi-supervised learning has been proposed to analyze when and why the unlabeled data is helpful [4].

Table 1
Classification of negative life events.

Label	Description	Example sentence
Family	Serious illness of a family member; son or daughter leaving home	<i>I am very worried about my children's health</i>
Love	Spouse/mate engaged in infidelity; broke up with a boyfriend or girlfriend	<i>I broke up with my dear but cruel boyfriend recently</i>
School	Examination failed or grade dropped; unable to enter/stay in school	<i>I hate to go to school because my teacher always blames me</i>
Work	Laid off or fired from a job; demotion and salary reduction	<i>I lost my job in this economic recession a few months ago</i>
Social	Substantial conflicts with a friend; difficulties in social activities	<i>I argued with my best friend and was upset</i>

Following the idea of semi-supervised learning, this study develops a weakly supervised framework by combining a supervised corpus-based method (association rule mining) and an unsupervised method, the *distributional semantic model* [18,30,50]. Instead of classifying the unlabeled data to augment the labeled data as used in the methods presented above, this study aims to acquire more association language patterns from the unlabeled data to augment the seed patterns generated from the labeled data. That is, the proposed framework requires only a small corpus of labeled sentences to generate a set of seed patterns using the supervised association rule mining. The unsupervised distributional semantic model is then performed to discover more language patterns semantically similar to the seed patterns from unlabeled web forum posts.

2. Methods

2.1. Dataset

2.1.1. Unlabeled corpus

The unlabeled corpus was a collection of forum posts collected from the two websites, John Tung Foundation (<http://www.jtf.org.tw>) and PsychPark (<http://www.psychpark.org>), a virtual psychiatric clinic, maintained by a group of volunteer professionals belonging to the Taiwan Association of Mental Health Informatics [3,28]. The forum posts from both websites were numbered according to the post time. We collected the 3500 latest forum posts from John Tung Foundation and the 1500 latest from PsychPark, which gave a total of 5000 forum posts in the unlabeled corpus.

2.1.2. Labeled corpus

The data to be labeled was collected from the Internet-based Self-Assessment Program for Depression (ISP-D) [29] database of PsychPark. The ISP-D comprises a potential maximum of 24 questions for a complete assessment, in which the second question asks users to list their negative life events or life stressors. Obviously, responses to this question were relatively clean compared with the forum posts because most responses contained negative life events. Such relatively clean data were suitable to be used to create the labeled corpus for seed pattern generation. We thus extracted sentences for annotation from the responses to the second question of the ISP-D database. Each sentence was labeled by two graduate students with one of the five types of negative life events described in Table 1. Disagreements between the two annotators were resolved by an adjudicator who was an experienced psychiatrist. Finally, a total of 2856 labeled sentences were obtained for the labeled corpus. The agreement of the two annotators was 87.8%. Table 2 presents the statistics of the labeled corpus. Table 3 shows the breakdown of the distribution of sentence types in the labeled corpus.

2.2. Overview of the system framework

Fig. 2 shows the overall framework of association language pattern mining. First, the association rule mining algorithm was adopted to mine a set of seed patterns by incrementally associating

Table 3

Distribution of sentence types in the labeled corpus.

Sentence type	% in corpus
Family	28.8
Love	22.8
School	13.3
Work	14.3
Social	20.8

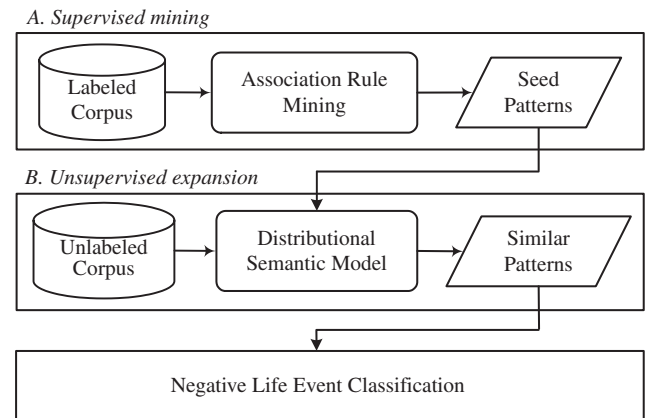


Fig. 2. Framework of association language pattern mining.

frequently co-occurred words from the labeled corpus. For each seed pattern, the distributional semantic model was used to discover similar patterns from the unlabeled web corpus. This involved a computation of the similarity between a seed pattern and a set of candidate patterns generated from the web corpus. The distributional semantic model accomplished this by comparing the context distributions of two patterns. The context distribution of a pattern, which can be retrieved from the web corpus, represents the co-occurrence frequency of the pattern and each word appearing in its context. Based on this contextual representation, two patterns sharing more common contexts are more similar semantically. Once all the seed patterns were exhausted, the discovered patterns (including the seed patterns) were used as features to train classifiers for negative life event classification.

2.3. Association rule mining

The problem of language pattern acquisition can be converted into the problem of association rule mining, in which each sales transaction in a database can be considered as a sentence in the corpora, and each item in a transaction denotes a word in a sentence. An association language pattern is defined herein as a combination of multiple associated words, denoted by $\langle w_1, \dots, w_k \rangle$. Thus, the task of association rule mining is to mine the language patterns of frequently associated words from the training sentences. For this purpose, we adopted the Apriori algorithm [1,9,13,52] and modified it slightly to fit our application. The basic concept behind the Apriori algorithm is the recursive identification of frequent word sets; association language patterns are then generated from the frequent word sets. For simplicity, only the combinations of nouns and verbs were considered, and the length of the word set was restricted to at most 4 words, i.e., 2-word, 3-word and 4-word combinations. The detailed procedure is described next.

2.3.1. Find frequent word sets

A word set is frequent if it possesses a minimum level of support. The support of a word set is defined as the number of labeled sentences containing the word set. For instance, the support of a

Table 2

Statistics of the labeled corpus.

Total number of records	1762
Total number of sentences in the records	2856
Avg. number of sentences per record	1.62
Min. number of sentences per record	1
Max. number of sentences per record	18

two-word set $\{w_i, w_j\}$ denotes the number of labeled sentences containing the word pair (w_i, w_j) . The frequent k -word sets are discovered from $(k-1)$ -word sets. First, the support of each word, i.e., the word frequency, in the labeled corpus was counted. The set of frequent one-word sets, denoted as L_1 , was then generated by choosing the words with a minimum support level. To calculate L_k , the following two-step process was performed iteratively until no more frequent k -word sets were found.

2.3.1.1. Join step. A set of candidate k -word sets, denoted as C_k , was first generated by merging frequent word sets of L_{k-1} , in which only the word sets whose first $(k-2)$ words were identical could be merged.

2.3.1.2. Prune step. The support of each candidate word set in C_k was then counted to determine which candidate word sets were frequent. Finally, the candidate word sets with a support count greater than or equal to the minimum support were considered to form L_k . The candidate word sets with a subset that was not frequent were eliminated. Fig. 3 shows an example of generating L_k .

2.3.2. Generate association patterns from frequent word sets

Association language patterns can be generated via a confidence measure once the frequent word sets have been identified. The confidence of an association language pattern of k words is defined as the mutual information of the k words, as shown below:

$$\text{Conf}(\langle w_1, \dots, w_k \rangle) = MI(w_1, \dots, w_k) \\ = P(w_1, \dots, w_k) \log \frac{P(w_1, \dots, w_k)}{\prod_{i=1}^k P(w_i)}, \quad (1)$$

where $P(w_1, \dots, w_k)$ denotes the probability of the k words co-occurring in the labeled corpus and $P(w_i)$ denotes the probability of a single word occurring in the labeled corpus. Accordingly, each frequent word set in L_k was assigned a mutual information score. To generate

a set of association language patterns, all frequent word sets were sorted in descending order of their mutual information scores. The minimum confidence (a threshold percentage) was then applied to select the top N percent of frequent word sets as the seed patterns. This threshold (α) was determined empirically on a development set (Section 3). Fig. 3 (right-hand side) shows an example of generating the association language patterns from L_k .

2.4. Distributional semantic model

The distributional semantic model is used to measure the semantic relatedness of two words (or patterns) by comparing their context distributions. Two words (or patterns) sharing more common contexts are more similar semantically. Previous research has shown that contextual information is useful for measuring word similarity [18,30,50,53]. For instance, consider the three words “boss”, “chief” and “flower” as an example. Words such as “stress”, “colleague”, and “company” often occur in the context of both “boss” and “chief”, but seldom occur in the context of “flower”. Hence, the words “boss” and “chief” are more similar because they have quite similar contexts. This study extends this notion to measure the similarity of two patterns so that the seed patterns derived from the previous section can be expanded by acquiring additional similar patterns from the unlabeled web corpus. To accomplish this goal, the distributional semantic model requires (1) a *representation scheme* to represent the context distribution of a word (or a pattern); (2) a *similarity measure* to measure the similarity between two words (or patterns) based on the contextual representation; and (3) a procedure for the *expansion of seed patterns*.

2.4.1. Representation scheme

2.4.1.1. Representation of a single word. The distributional semantic model uses a high-dimensional vector to record the co-occurrence information of a word and its context words. For instance, the contexts of a word w_k in a sentence $W = c_1 \dots c_{k-1} w_k c_{k+1} \dots c_n$ are

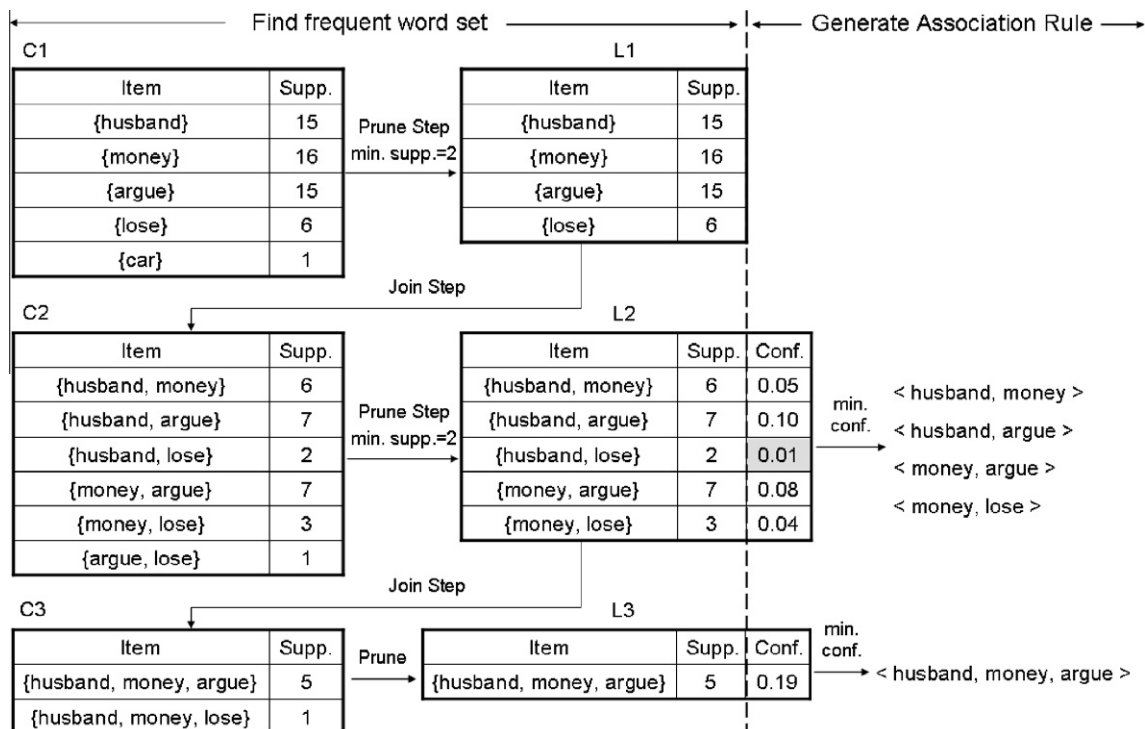


Fig. 3. Example of generating association language patterns.

$\{c_1, \dots, c_{k-1}, c_{k+1}, \dots, c_n\}$. Therefore, by considering all sentences containing w_k in the web corpus, the contexts of w_k are the words co-occurring with w_k in these sentences, which can be represented as:

$$v_{w_k} = \langle d_{w_k c_1}, d_{w_k c_2}, \dots, d_{w_k c_N} \rangle, \quad (2)$$

where $d_{w_k c_i}$ denotes the weight of the i th dimension of a vector, representing the strength of the association between w_k and its context word c_i , and N denotes the dimensionality of a vector, i.e., the number of distinct words appearing in the context of w_k in the corpus. The weight $d_{w_k c_i}$ is defined as:

$$d_{w_k c_i} = C(w_k, c_i), \quad (3)$$

where $C(w_k, c_i)$ is the number of times a word c_i appears in the context of w_k in the corpus.

To reduce the dimensionality of a context vector and measure the informativeness of a context word in each dimension, we applied the following rules. The sentences in the corpus are first segmented into word sequences. The distinct words in the word sequences (excluding punctuation marks) are considered as the dimension words to construct the vectors. Among the dimension words, the extremely infrequent words were considered to be noise and were discarded. Conversely, a high-frequency word generally received a higher weight, but this does not mean that it was informative because it could also appear in many other vectors. Therefore, the number of vectors in which a word appears should be considered when measuring its informativeness. In principle, a word appearing in more vectors carries less information to discriminate among the vectors. In this study, we adopted a weighting scheme analogous to TF-IDF [32,43] to re-weight the dimensions of a vector, as described in (4):

$$d_{w_k c_i} = d_{w_k c_i} * \log \frac{N(V)}{N(V_{c_i})}, \quad (4)$$

where $N(V)$ denotes the total number of vectors in the corpus and $N(V_{c_i})$ denotes the number of vectors with c_i as the dimension. The weight of each dimension can be further transformed into a probabilistic framework. That is:

$$d_{w_k c_i} \equiv P(c_i | w_k) = \frac{d_{w_k c_i}}{\sum_i d_{w_k c_i}}, \quad (5)$$

where $P(c_i | w_k)$ denotes the probability that c_i appears in the vector of w_k .

2.4.1.2. Representation of an association language pattern. Because an association language pattern consists of a set of words, it can be represented by combining the context vectors of its constituent words. A conceptual representation of context vector combination is shown in Fig. 4.

In Fig. 4, $alp_k = \langle w_1, \dots, w_n \rangle$ is an association language pattern with n constituent words. The dimensions (c_1, \dots, c_N) of each context vector are all distinct context words of the words in a pattern. For instance, let an association language pattern have two constituent words, w_1 and w_2 , where w_1 has three context words— c_1, c_2 ,

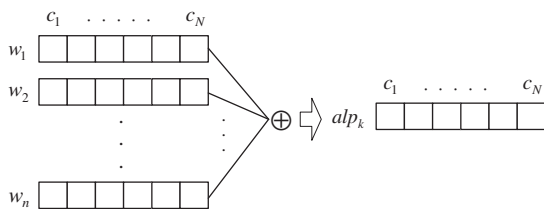


Fig. 4. The conceptual representation of context vector combination.

and c_3 —and w_2 has four context words— c_3, c_4, c_5 , and c_6 . The context vectors of both w_1 and w_2 can then be constructed using the six distinct context words (c_1, \dots, c_6) as the dimensions. If each dimension is a binary weight, then the context vector of w_1 and w_2 can be represented as $\langle 1, 1, 1, 0, 0, 0 \rangle$ and $\langle 0, 0, 1, 1, 1, 1 \rangle$, respectively. The representation of the context vector of alp_k can be formally defined as:

$$v_{alp_k} = \langle d_{alp_k c_1}, d_{alp_k c_2}, \dots, d_{alp_k c_N} \rangle, \quad (6)$$

where $d_{alp_k c_i}$ is the weight of the i th dimension of the context vector of an association language pattern, which is computed as the product of the weights of the i th dimension of its constituent words. That is:

$$d_{alp_k c_i} = \prod_{j=1}^n d_{w_j c_i} = \prod_{j=1}^n P(c_i | w_j). \quad (7)$$

According to the above product rule, the weight of the i th dimension of a pattern will be 0 if any constituent word in the pattern has an i th dimension of 0. The rationale behind using the product rule for context vector combination is that it can help filter noisy dimensions and retain the useful ones for pattern expansion.

2.4.2. Similarity measure

The previous section describes how each word (or pattern) in the corpus is associated with a vector representing its context distribution. Therefore, the similarity of two words (or patterns) can then be calculated by comparing their context distributions. As mentioned above, the weights of context vectors are transformed into a probabilistic framework (Eqs. (5) and (7)). Each context vector of a word (or a pattern) can thus be considered as a probabilistic distribution of its context words. Accordingly, the *Kullback–Leibler (KL) distance* [24] was adopted to calculate the distance between two probabilistic distributions. Let $v_{w_i} = \langle P(c_1 | w_i), \dots, P(c_N | w_i) \rangle$ and $v_{w_j} = \langle P(c_1 | w_j), \dots, P(c_N | w_j) \rangle$ be the context vectors (in probabilistic form) of the words w_i and w_j , respectively. The KL distance between these two vectors is defined as

$$D(v_{w_i} \| v_{w_j}) = \sum_{k=1}^N P(c_k | w_i) \log \frac{P(c_k | w_i)}{P(c_k | w_j)}, \quad (8)$$

where $D(\cdot \| \cdot)$ denotes the KL distance between two probabilistic distributions; $P(c_k | w)$ denotes the probabilistic weight of the k th dimension of the context vector of a word; and N denotes the dimensionality of a vector. The following divergence measure was adopted in the case of a symmetric distance:

$$Div(v_{w_i}, v_{w_j}) = D(v_{w_i} \| v_{w_j}) + D(v_{w_j} \| v_{w_i}). \quad (9)$$

In this way, the distance between two words can be calculated based on the KL divergence of their context vectors. That is;

$$Dist(w_i, w_j) = Div(v_{w_i}, v_{w_j}). \quad (10)$$

Therefore, the similarity between two words can be defined as

$$Sim(w_i, w_j) = \frac{1}{1 + Dist(w_i, w_j)}. \quad (11)$$

Eq. (11) shows that a smaller distance between two words indicates a greater similarity between them. Similarly, the similarity between two patterns can be calculated using Eqs. (8)–(10) by taking their context vectors as input. The similarity between two patterns can thus be defined as

$$Sim(alp_i, alp_j) = \frac{1}{1 + Dist(alp_i, alp_j)}. \quad (12)$$

2.4.3. Expansion of seed patterns

The seed patterns derived in the previous section were expanded by discovering additional similar patterns from the unlabeled web forum posts. This was accomplished by generating a set of candidate patterns for each seed pattern from the web corpus and then calculating the similarity of the candidate patterns to the seed pattern. As mentioned earlier, a pattern is composed of a set of nouns and verbs. Therefore, the candidate patterns for a seed pattern were all possible combinations of nouns and verbs in the corpus. However, discovering similar patterns from such a large dataset is inefficient. Additionally, not all words in the corpus are semantically related to the constituent words of a seed pattern. It is not necessary to include combinations of unrelated words because they are unlikely to be similar to the seed pattern. As a result, the number of candidate patterns for a seed pattern can be reduced by excluding the words dissimilar to the constituent words of the seed pattern. These steps are described below.

2.4.3.1. Candidate pattern generation. Given a seed pattern, the first step was to select a set of words similar to its constituent words. For each constituent word, the word-level similarity measure described in (11) was applied to evaluate all words in the corpus with the same part-of-speech as the constituent word. Only the words with a similarity score greater than the average of all words were retained for candidate pattern generation. The candidate patterns for a seed pattern were all possible combinations of the retained words.

2.4.3.2. Pattern expansion. Once the candidate patterns for a seed pattern were generated, they were all represented using the vector combination scheme. The similarity between each candidate pattern and the seed pattern was then calculated using the pattern-level similarity measure described in (12). Finally, these candidate patterns were ranked in descending order according to their similarity scores. Because not all candidate patterns contribute to the classification task, a threshold β was applied to select the top N percent of candidate patterns for classification. This threshold was determined empirically by maximizing the classification performance.

3. Results

3.1. Experiment setup

The labeled corpus (Section 2.1) was split into a training set, a development set, and a test set with an 8:1:1 ratio. The training set was used to generate seed patterns, the development set was used to optimize the thresholds for seed pattern generation (α) and expansion (β), and the optimal setting was used on the test set to evaluate the performance of negative life event classification. This experiment used 10-fold cross-validation for evaluation.

3.1.1. Features and classifiers

This experiment used the following feature set to train three different classifiers, including Support Vector Machine (SVM), C4.5, Naïve Bayes (NB), and Tree Augmented Naïve Bayes (TAN), which were provided by Weka Package [51].

- *Bag-of-words (BOW)*: Each single word in sentences.
- *Association language patterns (ALP)*: The seed patterns generated from the labeled corpus using association rule mining.
- *Web expansion (Web)*: The patterns expanded from the unlabeled web corpus using the distributional semantic model with the input of the seed patterns.
- *Ontology expansion (Onto)*: Another possible method to expand the seed patterns is the use of synonyms and hypernymy–hyponymy relations defined in a lexical ontology such as WordNet

(English) [14] and EuroWordNet (Multilingual) [42]. For example, the pattern <boss, conflict> can be expanded as <chief, conflict> because the words *boss* and *chief* are synonyms. This experiment used the HowNet to expand the seed patterns by mapping their constituent words into synonyms.

Each classifier was implemented using four different levels of features—namely BOW, BOW + ALP, BOW + ALP + Web, and BOW + ALP + Onto—to examine the effectiveness of association language patterns and the unsupervised pattern expansion for the classification task. For instance, BOW versus others examined the effectiveness of association language patterns, BOW + ALP versus BOW + ALP + Web or BOW + ALP + Onto examined the effectiveness of pattern expansion techniques, and BOW + ALP + Web versus BOW + ALP + Onto further compared the two expansion techniques: expansion from the web corpus and from a lexical ontology.

3.1.2. Evaluation metric

The classification performance was measured by the *accuracy*, i.e., the number of correctly classified sentences divided by the total number of test sentences.

3.2. Evaluation of threshold selection

The proposed framework involved two thresholds, α and β . The threshold α was used to control the number of seed patterns generated from the labeled corpus. The threshold β was applied in the later stage to control the number of patterns expanded from the web corpus. The best setting of these two thresholds was tuned for each individual classifier with different feature sets by maximizing the classification accuracy for the development set. This section uses Naïve Bayes (NB) as the example classifier to describe the threshold selection procedure. Fig. 5 shows the accuracy of NB against different settings of the threshold α .

When using association language patterns as features (BOW + ALP), the accuracy increased with increasing threshold values up to 0.6, indicating that the top 60% of discovered patterns contained more useful patterns for classification. By contrast, the accuracy decreased when the threshold value was above 0.6, indicating that the remaining 40% contained more noisy patterns that may increase the ambiguity in classification. When using the ontology expansion approach (BOW + ALP + Onto), both the number and diversity of discovered patterns increased. Therefore, the accuracy was improved, and the optimal accuracy was achieved at 0.5. However, the accuracy dropped significantly when the threshold value

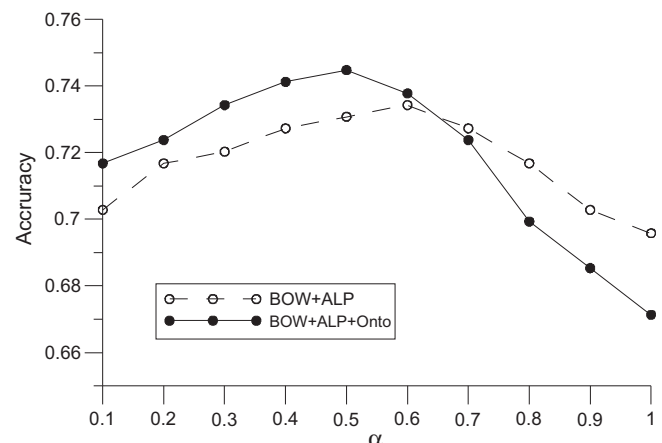


Fig. 5. Threshold selection (α).

was above 0.5. This finding indicates that expansion of noisy patterns may produce more noisy patterns and thus decrease performance. The total numbers of patterns used to build BOW + ALP and BOW + ALP + Onto were 638 and 927, respectively.

When using the web expansion approach (BOW + ALP + Web), both α and β should be considered in the tuning process. In principle, a higher α means that noisy patterns are more likely to be selected as seed patterns for expansion, thus generating more noisy patterns for classification. On the other hand, if α is too small, many useful patterns may not be discovered. Similarly, a higher β will also introduce more noisy patterns in the expansion. Fig. 6 shows the accuracy of NB against different settings of α and β .

With the increase of the threshold α (0.1, 0.3 and 0.5), more useful patterns were selected as the seed patterns for expansion, thus the performance increased accordingly with an appropriate expansion threshold β . For instance, the best settings of β for $\alpha = 0.1$, 0.3 and 0.5 were 0.4, 0.3, and 0.2, respectively. When β exceeded the best setting, the performance decreased rapidly because most of the expanded patterns at lower ranks were noisy patterns. This phenomenon deteriorated when α exceeded 0.5 because more noisy patterns were selected as the seed patterns for expansion, yielding many more noisy patterns expanded from the web corpus. The best settings of the thresholds for BOW + ALP + Web for the NB classifier were $\alpha = 0.5$ and $\beta = 0.2$, which means that the top 50% of patterns produced by association rule mining were selected as seed patterns for expansion, and the top 20% of patterns expanded from the web corpus were selected for classifier training. The total number of patterns used to build BOW + ALP + Web was 1731.

3.3. Results of classification performance

The results of each classifier were obtained from the test set, using the best setting of the thresholds optimized in the previous section. Table 4 shows the comparative results of different classifiers with different levels of features. The paired, two-tailed *t*-test was used to determine whether the performance difference was statistically significant.

Compared with the baseline feature (BOW), the use of association language patterns (BOW + ALP) improved the accuracy of NB, C4.5, TAN, and SVM by 3.7%, 1.6%, 3.5%, and 2.2%, respectively, and achieved an average improvement of 2.7%. Additionally, the use of the two unsupervised expansion methods, BOW + ALP + Onto and BOW + ALP + Web, further improved the accuracy by 1.6% and

4.8% on average, respectively, compared with BOW + ALP. This finding indicates that association language patterns are significant features for negative life event classification. The reasons behind the improvement are described as follows. The baseline feature BOW treats each word independently, without considering the relationships of words in sentences. Classifiers trained with BOW alone thus tend to be ambiguous regarding sentences that contain the same words but are categorized as describing different negative life events. The association language patterns can instead capture both the local and long-distance dependencies of words in sentences, which help to increase the classifiers' ability to distinguish among different negative life events, thus yielding a higher accuracy. Additionally, the unsupervised expansion methods can increase both the number and diversity of the patterns by mining the unlabeled web corpus, further improving the performance of the supervised mining method (BOW + ALP). Comparing the information sources used for pattern expansion, BOW + ALP + Web achieved a higher level of performance than did BOW + ALP + Onto. The possible reasons for this are twofold. First, the ontology-based approach used only synonym information for pattern expansion and thus tended to miss useful patterns consisting of near-synonyms or other related words. Second, a lexical ontology is a static knowledge resource, which might not reflect the dynamic characteristics of language.

For a more detailed analysis, the class-by-class performance measured by the area under the ROC curve (AUC) [12,17] was retrieved from the involved classifiers with BOW + ALP + Web, as shown in Table 5. The results show that the classes <Family> and <Love> had relatively better performance than the other classes. Another observation is that NB, TAN, and SVM had similar average AUC scores, and all of them achieved significant higher performance than C4.5.

Table 6 shows the confusion matrix for the SVM classifier with BOW + ALP + Web. In <Family>, there was a total of 118 misclassified test examples, where 43% (51/118) of them were classified into <Love>. Similarly, 65% (75/115) of the misclassified examples in <Love> were classified into <Family>. This finding indicates that the two classes <Family> and <Love> are ambiguous with each other. Another observation is that no test example in <School> was misclassified into <Work>, and vice versa. This indicates that the two classes <School> and <Work> were easily separable.

3.4. Comparison with self-training

In BOW + ALP, the model was trained with the seed patterns generated from the labeled corpus using association rule mining. BOW + ALP + Web further expanded the seed patterns from the unlabeled web corpus using the unsupervised distributional semantic model. Self-training is also a possible method to expand the seed patterns by automatically and iteratively classifying the unlabeled web corpus to augment the labeled corpus. More useful patterns can then be discovered from the augmented labeled corpus to improve the classification performance. Below are the detailed steps of self-training for pattern expansion. A base classifier was first built using BOW + ALP. The base classifier was then used to classify the unlabeled examples, and the most confidently classified examples were added to the labeled corpus. The association rule mining was then used to generate a new feature set (patterns) from the augmented labeled corpus. The base classifier was re-trained with the new feature set, and the process is iterated. In each iteration, examples were considered to be confidently classified if their scores (output by each classifier) were greater than 0.8. Of the confidently classified examples, the top 150 (30 for each class) most confidently classified examples were added to augment the labeled corpus. This process was stopped until all confidently classified examples have been added. Table 7 shows the compara-

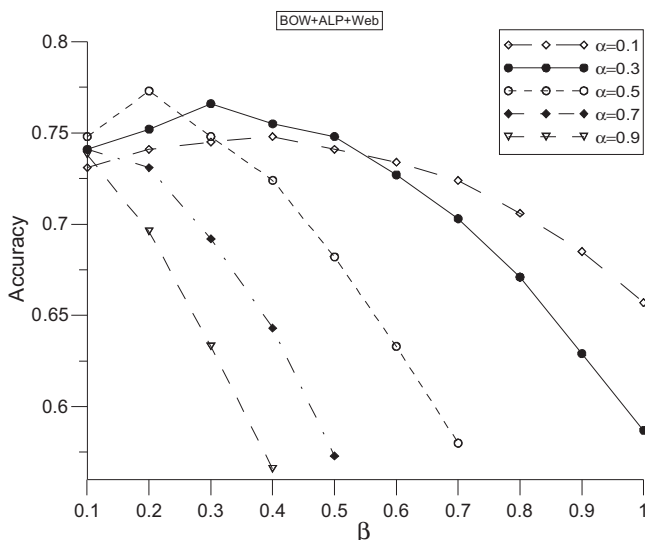


Fig. 6. Threshold selection (β).

Table 4

Comparative results of different classifiers with different levels of features.

	Baseline	Supervised mining	Unsupervised expansion	
	BOW	BOW + ALP	BOW + ALP + Onto	BOW + ALP + Web
NB	0.702 ± 0.033	0.728 ± 0.034 ^a	0.746 ± 0.036	0.772 ± 0.028 ^b
C4.5	0.734 ± 0.030	0.746 ± 0.035	0.755 ± 0.032	0.776 ± 0.027 ^b
TAN	0.723 ± 0.032	0.748 ± 0.031 ^a	0.761 ± 0.027	0.793 ± 0.030 ^b
SVM	0.778 ± 0.028	0.795 ± 0.026	0.803 ± 0.025	0.819 ± 0.021 ^b
Avg.	0.734 ± 0.041	0.754 ± 0.039 ^a	0.766 ± 0.036	0.790 ± 0.032 ^b

^a BOW + ALP vs BOW significantly different ($p < 0.05$).^b BOW + ALP + Web vs BOW + ALP significantly different ($p < 0.05$).**Table 5**

Class-by-class performance (AUC).

	NB	TAN	C4.5	SVM
Family	0.919 ± 0.038	0.910 ± 0.036	0.866 ± 0.045	0.922 ± 0.026
Love	0.905 ± 0.047	0.931 ± 0.028	0.879 ± 0.041	0.901 ± 0.027
School	0.878 ± 0.066	0.898 ± 0.052	0.851 ± 0.081	0.909 ± 0.039
Work	0.857 ± 0.073	0.889 ± 0.078	0.807 ± 0.097	0.863 ± 0.058
Social	0.887 ± 0.062	0.884 ± 0.055	0.843 ± 0.076	0.875 ± 0.061
Avg.	0.894 ± 0.034 ^a	0.906 ± 0.031 ^a	0.854 ± 0.038	0.897 ± 0.025 ^a

^a Classifiers vs C4.5 significantly different ($p < 0.05$).**Table 6**

Confusion matrix for SVM with BOW + ALP + Web.

	Family	Love	School	Work	Social
Family	705	51	21	16	30
Love	75	536	7	11	22
School	26	19	312	5	18
Work	45	25	3	308	27
Social	36	30	28	21	479

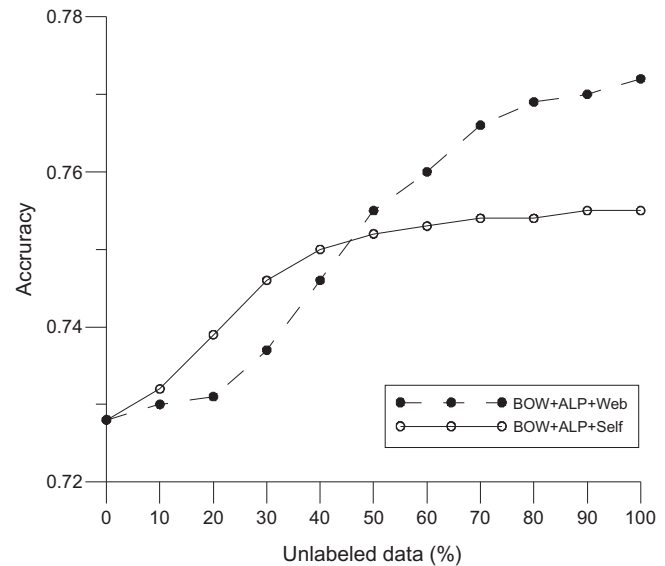
tive results of different classifiers built with self-training (BOW + ALP + Self) and BOW + ALP + Web. The results show that BOW + ALP + Web yielded higher performance than BOW + ALP + Self.

For a more detailed analysis, Fig. 7 uses NB as the example classifier to show the comparative results of using different sizes of unlabeled data for seed pattern expansion. For self-training, more patterns can be generated from the iteratively augmented labeled corpus using association rule mining. However, association rule mining uses the mutual information (Section 2.3.2) to discover word associations within the sentence boundary, which may generate more common patterns when the sentences in the corpus share too many common words. That is, the diversity of the discovered patterns may not increase when more data was used for expansion. As indicated in Fig. 7, the performance of BOW + ALP + Self increased rapidly when less than 40% unlabeled data was used because both quantity and diversity of the discovered patterns increased in this stage. When more than 40% unlabeled data was used, the performance of BOW + ALP + Self was similar because less diverse patterns were discovered in this stage. The distributional semantic model can relax the limitation of association rule

Table 7

Comparative results of self-training and distributional semantic model.

	BOW + ALP + Self	BOW + ALP + Web
NB	0.753 ± 0.026	0.772 ± 0.028
C4.5	0.764 ± 0.029	0.776 ± 0.027
TAN	0.772 ± 0.031	0.793 ± 0.030
SVM	0.808 ± 0.023	0.819 ± 0.021
Avg.	0.774 ± 0.033	0.790 ± 0.032 [*]

^{*} Significantly different ($p < 0.05$).**Fig. 7.** Comparative results of using different sizes of unlabeled data for seed pattern expansion.

mining by discovering word associations across sentences using the context distributions retrieved from the whole unlabeled corpus (Section 2.4). This is reason why the performance of BOW + ALP + Web kept increasing as more unlabeled data was added.

3.5. Evaluation on the size of the labeled and unlabeled datasets

The size of the datasets used for seed pattern generation (labeled corpus) and expansion (unlabeled web corpus) also affected the classification performance. To investigate this effect, this experiment first randomly divided both the labeled and unlabeled corpora into five equal folds. Different dataset sizes could then be used to build the classifiers. Fig. 8 uses NB as the example classifier to show the accuracy against different dataset sizes. Table 8 shows the accuracy of some of the data points in Fig. 8.

As expected, the performance increased when more labeled data was used. However, this also increased the annotation costs. The use of unlabeled data is an alternative approach to improving the performance. As indicated in Fig. 8, the performance shows a stable tendency to increase as more unlabeled data was used. For instance, when using 60% of the labeled data combined with the unlabeled data, the accuracy kept increasing as more unlabeled data was added and finally became higher (0.752, Table 8) than when using 100% labeled data alone (0.728, Table 8). This finding indicates that the use of an unlabeled web corpus for pattern expansion can not only improve the level of performance but also reduce the reliance on labeled corpora.

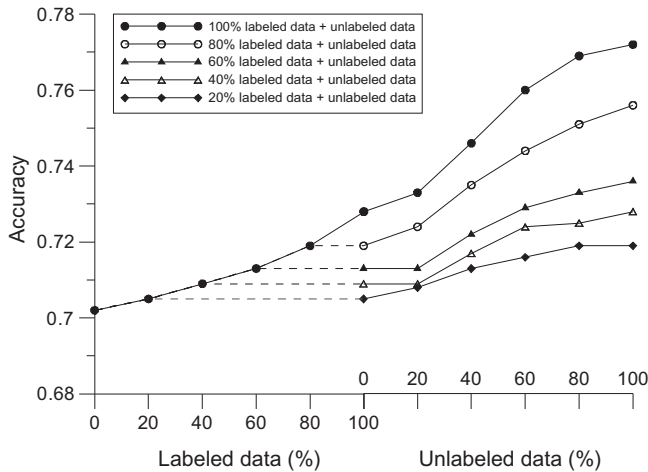


Fig. 8. Performance against different sizes of the labeled and unlabeled datasets.

Table 8

Accuracy of using different portions of the labeled data combined with 100% of the unlabeled data.

	Accuracy
20% labeled data + 100% unlabeled data	0.719
40% labeled data + 100% unlabeled data	0.725
60% labeled data + 100% unlabeled data	0.736
80% labeled data + 100% unlabeled data	0.756
100% labeled data + 100% unlabeled data	0.772
100% labeled data	0.728

4. Conclusion

In this study, we presented a framework that combines a supervised data mining algorithm and an unsupervised expansion method to acquire association language patterns for negative life event classification. The supervised data mining algorithm—association rule mining—was used to generate a set of seed patterns from a labeled corpus. The unsupervised expansion method—using the distributional semantic model—was then performed to discover more patterns similar to the seed patterns from an unlabeled web corpus. The experimental results show that association language patterns are promising features for classification tasks because they can capture word relationships in sentences. The unsupervised expansion method can further improve classification performance because it can increase both the number and diversity of discovered patterns. It also enables the proposed framework to bootstrap using a small amount of labeled data, thus reducing the reliance of the classification process on the availability of a large, labeled corpus.

Our future work will be devoted to addressing some of the limitations of this work. First, the current approach will be extended with multi-category classification technologies to classify the sentences containing multiple or no negative life events. Second, in addition to nouns and verbs, adjectives such as “bad”, “horrible”, and “difficult” will be considered for inclusion in the association language pattern mining process.

Acknowledgment

This work was supported by the National Science Council, Taiwan, ROC, under Grant No. NSC98-2221-E-155-052. The authors would like to thank the anonymous reviewers and editors for their constructive comments.

References

- [1] Agrawal R, Srikant R. Fast algorithms for mining association rules. In: Proc int'l conf very large data bases (VLDB); 1994. p. 487–99.
- [2] Atkinson J, Rivas A. Discovering novel causal patterns from biomedical natural-language texts using bayesian nets. *IEEE Trans Inf Technol Biomed* 2008;12(6):714–22.
- [3] Bai YM, Lin CC, Chen JY, Liu WC. Virtual psychiatric clinics. *Am J Psychiat* 2001;158(7):1160–1.
- [4] Balcan MF, Blum A. A discriminative model for semi-supervised learning. *J. ACM* 2010;57(3). article 19.
- [5] Blum A, Mitchell T. Combining labeled and unlabeled data with co-training. In: Proc of the eleventh annual conference on computational learning theory; 1998. p. 92–100.
- [6] Brefeld U, Scheffer T. Co-EM support vector learning. In: Proc of ICML-04; 2004. p. 121–8.
- [7] Brostedt EM, Pedersen NL. Stressful life events and affective illness. *Acta Psychiatr Scand* 2003;107(3):208–15.
- [8] Chapelle O, Schölkopf B, Zien A. Semi-supervised learning. Cambridge, MA: MIT Press; 2006.
- [9] Chien JT. Association pattern language modeling. *IEEE Trans Audio Speech Lang Process* 2006;14(5):1719–28.
- [10] Conway M, Doan S, Kawazoe A, Collier N. Classifying disease outbreak reports using *n*-grams and semantic features. *Int J Med Inform* 2009;78(12):e47–58.
- [11] Exarchos TP, Papaloukas C, Lampros C, Fotiadis DI. Mining sequential patterns for protein fold recognition. *J Biomed Inform* 2008;41(1):165–79.
- [12] Fawcett T. An introduction to ROC analysis. *Pattern Recognit Lett* 2006;27(8):861–74.
- [13] Fayyad UM, Piatetsky-Shapiro G, Smyth P, Uthurusamy R. Advances in knowledge discovery & data mining. Cambridge (MA): MIT Press; 1996.
- [14] Fellbaum C. WordNet: an electronic lexical database. Cambridge (MA): MIT Press; 1998.
- [15] Friedman N, Geiger D, Goldszmidt M. Bayesian network classifiers. *Mach Learn* 1997;29(2–3):131–63.
- [16] Guz U, Cuendet S, Hakkani-Tür D, Tur G. Multi-view semi-supervised learning for dialog act segmentation of speech. *IEEE Trans Audio Speech Lang Process* 2010;18(2):320–9.
- [17] Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* 1982;143(1):29–36.
- [18] Harris Z. Distributional structure. *Word* 1954;10(2–3):146–62.
- [19] Huang M, Zhu X, Li M. A hybrid method for relation extraction from biomedical literature. *Int J Med Inform* 2006;75(6):443–55.
- [20] Joachims T. Transductive inference for text classification using support vector machines. In: Proc of ICML-99; 1999. p. 200–9.
- [21] Khoo A, Marom Y, Albrecht D. Experiments with sentence classification. In: Proc of australasian language technology workshop; 2006. p. 18–25.
- [22] Kim H, Chen SS. Associative naïve Bayes classifier: automated linking of gene ontology to medline documents. *Pattern Recognit* 2009;42(9):1777–85.
- [23] Kononenko I. Semi-naïve Bayesian classifier. In: Proc of sixth european working session on learning; 1991. p. 206–19.
- [24] Kullback S. Information theory and statistics. New York: John-Wiley & Sons; 1959.
- [25] Lan M, Tan CL, Su J. Feature generation and representations for protein–protein interaction classification. *J Biomed Inform* 2009;42(5):866–72.
- [26] Lehnert W, Cardie C, Fisher D, McCarthy J, Riloff E, Soderland S. University of massachusetts: description of the CIRCUS system used for MUC-4. In: Proc fourth message understanding conference (MUC-4); 1992. p. 282–8.
- [27] Lertnate V, Theeramunkong T. Multidimensional text classification for drug information. *IEEE Trans Inf Technol Biomed* 2004;8(3):306–12.
- [28] Lin CC, Bai YM, Chen JY. Reliability of information provided by patients of a virtual psychiatric clinic. *Psychiat Serv* 2003;54(8):1167–8.
- [29] Lin CC, Li YC, Bai YM, Tsai SJ, Liu CY, Hsiao MC et al. The validity of an Internet-based Self-assessment Program for Depression. In: Proc of AMIA; 2003.
- [30] Lin D. Automatic retrieval and clustering of similar words. In: Proc of ACL-98; 1998. p. 768–74.
- [31] McClosky D, Charniak E. Self-training for biomedical parsing. In: Proc of ACL-08; 2008. p. 101–4.
- [32] Montañés E, Díaz I, Ranilla J, Combarro EF, Fernández J. Elena scoring and selecting terms for text categorization. *IEEE Intell Syst* 2005;20(3):40–7.
- [33] Murata M, Mitsumori T, Doi K. Analysis and improved recognition of protein names using transductive SVM. *J Comput* 2008;3(1):51–62.
- [34] Muslea I. Extraction patterns for information extraction tasks: a survey. In: Proc AAAI workshop on machine learning for information extraction; 1999. p. 1–6.
- [35] Mykowiecka A, Marciniak M, Kuśc A. Rule-based information extraction from patients' clinical data. *J Biomed Inform* 2009;42(5):923–36.
- [36] Naughton M, Stokes N, Carthy J. Investigating statistical techniques for sentence-level event classification. In: Proc of COLING-08; 2008. p. 617–24.
- [37] Nigam K, Ghani R. Analyzing the effectiveness and applicability of co-training. In: Proc of CIKM-00; 2000. p. 86–93.
- [38] Nigam K, McCallum AK, Thrun S, Mitchell T. Text classification from labeled and unlabeled documents using EM. *Mach Learn* 2000; 39(2–3):103–34.

- [39] Pagano ME, Skodol AE, Stout RL, Shea MT, Yen S, Grilo CM, Sanislow CA, Bender DS, McGlashan TH, Zanarini MC, Gunderson JG. Stressful life events as predictors of functioning: findings from the collaborative longitudinal personality disorders study. *Acta Psychiatr Scand* 2004;110(6):421–9.
- [40] Paradis F, Nie JY. Contextual feature selection for text classification. *Inf Process Manage* 2007;43(2):344–52.
- [41] Paykel ES. Life events and affective disorders. *Acta Psychiatr Scand* 2003;108(Suppl. 418):61–6.
- [42] Rodríguez H, Climent S, Vossen P, Bloksma L, Peters W, Alonge A, Bertagna F, Roventint A. The top-down strategy for building EuroWordNet: vocabulary coverage, base concepts and top ontology. *Comput Humanities* 1998;32:117–59.
- [43] Salton G, Buckley C. Term-weighting approaches in automatic text retrieval. *Inf Process Manage* 1988;24(5):513–23.
- [44] Scudder HJ. Probability of error of some adaptive pattern-recognition machines. *IEEE Trans Inf Theory* 1965;11(3):363–71.
- [45] Sebastiani F. Machine learning in automated text categorization. *ACM Comput Surv* 2002;34(1):1–47.
- [46] Tai YM, Chiu HW. Comorbidity study of ADHD: applying association rule mining (ARM) to National Health Insurance Database of Taiwan. *Int J Med Inform* 2009;78(12):e75–83.
- [47] Talukdar PP, Pereira F. Experiments in graph-based semi-supervised learning methods for class-instance acquisition. In: *Proc of ACL-10*; 2010. p. 1473–81.
- [48] Tan CM, Wang YF, Lee CD. The use of bigrams to enhance text categorization. *Inf Process Manage* 2002;38(4):529–46.
- [49] Wang J, Shen X, Pan W. On transductive support vector machines. In: Verducci J, Shen X, Lafferty J, editors. *Prediction and discovery*. American Mathematical Society; 2007.
- [50] Weeds J, Weir D, McCarthy D. Characterising measures of lexical distributional similarity. In: *Proc of COLING-04*; 2004.
- [51] Witten IH, Frank E. *Data mining: practical machine learning tools and techniques*. 2nd ed. San Francisco: Morgan Kaufmann; 2005.
- [52] Wu CH, Chuang ZJ, Lin YC. Emotion recognition from text using semantic labels and separable mixture models. *ACM Trans Asian Lang Inform Process* 2006;5(2):165–82.
- [53] Yeh JF, Wu CH, Yu LC, Lai YS. Extended probabilistic HAL with close temporal association for psychiatric consultation query retrieval. *ACM Trans Inf Syst* 2008;27(1). article 4.
- [54] Yu LC, Wu CH, Chang RY, Liu CH, Hovy EH. Annotation and verification of sense pools in OntoNotes. *Inf Process Manage* 2010;46(4):436–47.
- [55] Yu LC, Wu CH, Jang FL. Psychiatric document retrieval using a discourse-aware model. *Artif Intell* 2009;173(7–8):817–29.
- [56] Zanzotto FM, Pennacchiotti M. Expanding textual entailment corpora from Wikipedia using co-training. In: *Proc of the second workshop on collaboratively constructed semantic resources at coling-10*; 2010. p. 28–36.
- [57] Zhu X. *Semi-supervised learning literature survey*. computer sciences TR 1530. University of Wisconsin-Madison; 2008.
- [58] Zhu X, Ghahramani Z, Lafferty J. Semi-supervised learning using gaussian fields and harmonic functions. In: *Proc. of ICML-03*; 2003. p. 912–9.